Nucleic Acids Research Advance Access published October 11, 2007

Nucleic Acids Research, 2007, 1–4
doi:10.1093/nar/gkm798

The Generation Challenge Programme comparative plant stress-responsive gene catalogue

Samart Wanchana[1], Supat Thongjuea[1,2], Victor Jun Ulat[1], Mylah Anacleto[1], Ramil Mauleon[1], Matthieu Conte[3], Mathieu Rouard[4], Manuel Ruiz[3], Nandini Krishnamurthy[5], Kimmen Sjolander[5], Theo van Hintum[6] and Richard M. Bruskiewich[1,*]

[1]Crop Research Informatics Laboratory – International Rice Research Institute (IRRI), DAPO Box 7777, Metro Manila, Philippines, [2]National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, Phahonyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand, [3]Centre International de Recherche Agronomique pour le Développement (CIRAD), Avenue Agropolis, 34398 Montpellier, Cedex 5, France, [4]Bioversity International, Via dei Tre Denari 472/a, 00057 Maccarese, Rome, Italy, [5]Department of Bioengineering, University of California, 473 Evans Hall #1762, Berkeley, CA 94720, USA and [6]Wageningen Universiteit & Researchcentrum (WUR), 6700 HB Wageningen, Germany

Received August 14, 2007; Revised September 16, 2007; Accepted September 17, 2007

ABSTRACT

The Generation Challenge Programme (GCP; www.generationcp.org) has developed an online resource documenting stress-responsive genes comparatively across plant species. This public resource is a compendium of protein families, phylogenetic trees, multiple sequence alignments (MSA) and associated experimental evidence. The central objective of this resource is to elucidate orthologous and paralogous relationships between plant genes that may be involved in response to environmental stress, mainly abiotic stresses such as water deficit ('drought'). The web-based graphical user interface (GUI) of the resource includes query and visualization tools that allow diverse searches and browsing of the underlying project database. The web interface can be accessed at http://dayhoff.generationcp.org.

INTRODUCTION

Comparative biology provides valuable insights into organismal function and evolution, highlighting the divergence and conservation of gene families and biological processes. In order to cross-reference genes from one species to other related species, accurate predictions of orthologous and paralogous relationships are necessary. Such cross-referencing potentially permits researchers to infer the molecular functions of genes lacking such annotation from experiments in other, better-characterized organisms. Paralogous genes arising from ancient duplication events are likely to have diverged in function, whereas orthologous genes with common ancestry separated only by speciation are more likely to retain identical or highly similar function over evolutionary time (1,2). Such orthologous and paralogous gene loci almost invariably share some common molecular characteristics; thus, important inferences of function may be possible once these relationships are clearly defined.

The Generation Challenge Programme (GCP; www.generationcp.org) is a global crop research consortium striving to apply comparative genomics and molecular analysis to plant genetic resources to enhance efforts in plant breeding for plant stress tolerance. Clustering of orthologous genes across multiple crop species is a powerful strategy for the identification of stress-responsive gene loci and their corresponding alleles of high agronomic value, for application in breeding for stress tolerance.

To facilitate cross-species gene functional analysis, the GCP commissioned a project to assemble tools for the compilation and visualization of comparative information about stress-responsive genes. The result is an online resource, code-named Dayhoff, after Margaret Dayhoff, the famous early pioneer in comparative analysis of sequences.

Orthologues and paralogues of stress-responsive genes are presented by means of phylogenetic trees constructed using a phylogenomic inference method (3,4). The Dayhoff catalogue is expected to guide the bioinformatics analysis and interpretation of research results generated by comparative genomics experiments. For example,

*To whom correspondence should be addressed. Tel: +63 2 580 5600; Fax: +63 2 580 5699; Email: r.bruskiewich@cgiar.org

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

microarray data about drought stress obtained across diverse crop species will be analysed in a comparative manner to identify conserved gene expression profiles exhibited under similar stresses, in a similar fashion to experiments in other model species (5–7).

DATABASE CONSTRUCTION AND IMPLEMENTATION

Dayhoff is a MySQL database based mainly on the Chado schemata of the Generic Model Organism Database project (8) (www.gmod.org), with local enhancements where necessary, to store protein family information such as protein multiple sequence alignments (MSA), phylogenetic trees and supported stress evidence from experiments and the literature. The web interface uses GCP Java-based software technology (http://pantheon.generationcp.org) connected to third-party software such as ATV (9), Jalview (10) and BLAST (11) for analysing and viewing the query's results. The Dayhoff site is also cross-linked to a complementary GCP-funded comparative gene analysis resource called GreenPhyl. GreenPhyl provides comparative genomic analyses of *Arabidopsis thaliana* and *Oryza sativa* whole-genome assemblies and can be accessed directly at http://greenphyl.cirad.fr/cgi-bin/greenphyl.cgi.

DATA ANALYSIS AND CURATION

The core data set in Dayhoff consists of stress-related protein families characterized by a phylogenomic inference approach (4,12). The method has been shown to enable the highest accuracy in predicting protein molecular function (12), to avoid most false homology inference problems, and to distinguish between orthologous and paralogous genes (4). Phylogenetic trees representing protein families were constructed by the following steps. First, homologous sequences for each stress protein compiled from the literature were gathered by using the FlowerPower tool on the Berkeley Phylogenomics Group (BPG) web server (13), with Uniprot proteins (14) used as a database. FlowerPower uses iterative subfamily hidden Markov model (HMM) searches against PSI-BLAST-identified homologues and alignment analysis to discriminate between partial and global homologies (12). Then, MSAs of homologous proteins were constructed with the high-accuracy MSA program, MUSCLE v. 3.52 (15). After masking the alignments to remove columns with many gap characters, functional subfamilies were identified for each group using the SCI-PHY web server (12). SCI-PHY uses Bayesian and information-theoretic approaches to construct a hierarchical tree and cut tree into subtrees to identify functional subfamilies (12). The analysed trees were saved in the extended New Hampshire format (NHX) for display by the ATV program (9).

Stress-responsive genes to be analysed were compiled from available literature documenting genes analysed from diverse experimental sources (Supplementary Table 1). In the current version of Dayhoff, stress genes include those analysed from drought, salt, cold, ABA and GA stress experiments. Both up- and down-regulated genes under those stress types are available for *O. sativa* and *A. thaliana*. To overlay this experimental evidence on the gene family trees, BLASTP searches of candidate stress genes were performed against the database of Uniprot proteins used in phylogenetic tree construction. The BLAST results were limited into the ranks of parameter cutoff values as following: $\geq$80% to >95% similarity, E-value $<1e-20$ to $<1e-50$ and bit scores >50 to >1000.

USER INTERFACE

There are three main options for using the database: browsing protein families, query database by gene names or protein names and BLAST search against protein families (Figure 1).

Browsing protein families

The database can be used by browsing the entire set of stress protein families that have been constructed (Figure 1A). Users can select for browsing the database from the main drop-down menu. A list of protein families as well as links for phylogenetic trees and MSA are shown on the front page. Details about each protein family, for example, the list of Uniprot IDs, protein names, Gene Ontology (GO) terms and key publications for each protein obtained from Uniprot database (14), can be accessed through the family ID links (Figure 1B). Additional information can be displayed by selecting from the drop-down list. MSAs and the phylogenetic trees can be viewed by Jalview and ATV, respectively (Figure 1C and D). There are two choices for presenting the MSAs, by a whole family or users can select some proteins of interest to be aligned by checking the check boxes (Figure 1B). Hyperlinks to the Uniprot database and other online resources are also provided. Users can find stress evidence mapped to the matched protein(s) in the family owing to the BLASTP search results (Figure 1E). BlastP cutoff values for% identity, E-value and score are provided for filtering the BLASTP results. Users may need to change the default parameters in order to receive optimum results.

Query database

In the current version of Dayhoff, users can search the database by keywords within two fields of data type: Family name and Protein name (Figure 1G). By searching Family name, the matched family will be retrieved. Users can view more information through the family ID link as well as MSA and tree links. By searching Protein name, matched protein(s) will be listed together with Family ID link and some other information.

BLAST protein families

Users can submit a protein or DNA sequence in Fasta or raw format in order to BLAST the Dayhoff database as well as the GreenPhyl database (Figure 1H). Dayhoff is interconnected to the GreenPhyl database via a GCP-compliant BioMOBY (16) client web service. Users will

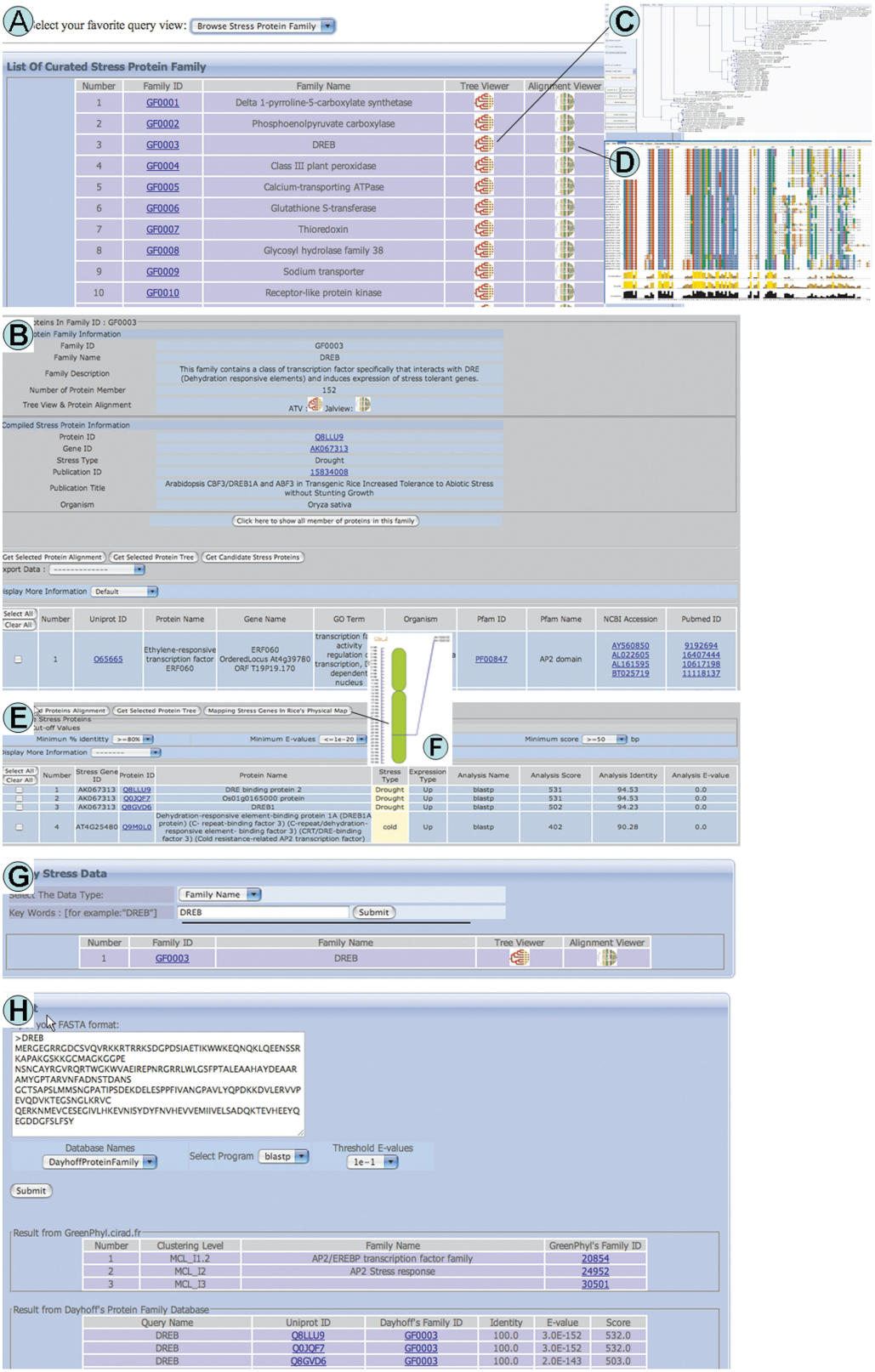

Figure 1. An example of browsing (**A**), querying (**G**) and BLAST searching the Dayhoff database (**H**). Phylogenetic tree and multiple sequence alignment of each protein family are displayed by ATV (**C**) and Jalview (**D**), respectively. Information on each protein family is shown in a new window (**B**) from the protein family ID links. Candidate stress proteins can be viewed in a new window when you toggle the *Get Candidate Stress Protein* button (B and **E**). A location of stress genes in the rice genome is drawn in the chromosome graphic (**F**). Dayhoff can be queried by protein names or family names (G). A protein sequence or nucleotide sequence can be submitted to BLAST the Dayhoff and GreenPhyl databases. The results of BLAST search are provided in both Dayhoff protein families and GreenPhyl classified families (H).

receive the results of best hits of protein family from both Dayhoff and GreenPhyl. The results will be provided with links to Dayhoff protein families and hyperlinks to classified families at the GreenPhyl web site.

FUTURE DIRECTIONS

Further integration of the comparative stress-responsive gene catalogue with the GCP platform software will enhance access to comparative gene data in a variety of bioinformatics analysis contexts. In particular, Dayhoff will be connected using GCP technology to a MAXD gene expression database, for direct integration into comparative microarray data analyses.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by Generation Challenge Programme.

Conflict of interest statement. None declared.

REFERENCES

1. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
2. Thornton,J.W. and DeSalle,R. (2000) Gene family evolution and homology: genomics meets phylogenetics. *Annu. Rev. Genomics Hum. Genet.*, **1**, 41–73.
3. Brown,D. and Sjolander,K. (2006) Functional classification using phylogenomic inference. *PLoS Computat. Biol.*, **2**, e77.
4. Sjolander,K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170–179.
5. Bergmann,S., Ihmels,J. and Barkai,N. (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.*, **2**, e9.
6. McCarroll,S.A., Murphy,C.T., Zou,S., Pletcher,S.D., Chin,C.-S., Jan,Y.N., Kenyon,C., Bargmann,C.I. and Li,H. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.*, **36**, 197–204.
7. Zhou,X. and Gibson,G. (2004) Cross-species comparison of genome-wide expression patterns. *Genome Biol.*, **5**, 232.
8. Mungall,C.J. and Emmert,D.B. (2007) The FlyBase C: a Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
9. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
10. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
11. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
12. Glanville,J.G., Kirshner,D., Krishnamurthy,N. and Sjolander,K. (2007) Berkeley Phylogenomics group web servers: resources for structural phylogenomic analysis. *Nucleic Acids Res.*, **35**, W27–W32.
13. Krishnamurthy,N., Brown,D. and Sjolander,K. (2007) FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evol. Biol.*, **7**, S12.
14. The UniProt C. (2007) The universal protein resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
15. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
16. Wilkinson,M., Schoof,H., Ernst,R. and Haase,D. (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The planet exemplar case. *Plant Physiol.*, **138**, 5–17.